CONVEX CATEGORIES AND ENTROPY: NOTES FOR UCONN'S S.I.G.M.A. SEMINAR

ARTHUR J. PARZYGNAT

Abstract. The concept of entropy arises in a plethora of disparate scenarios. To name a few, there are Boltzmann's definition in terms of statistical ensembles of particles, Shannon's information-theoretic notion of entropy, von Neumann's (quantum) entropy, Bekenstein's black hole entropy, Kolmogorov-Sinai's entropy for dynamical systems, and even Perelman's entropy in geometry which was an ingredient in Perelman's proof of the Poincare conjecture. This talk focuses on Shannon entropy from a categorical perspective and provides an overview of recent work of Baez, Fritz, and Leinster. The main point of their work is that although entropy is not characterized as a convex function on probability spaces, it is characterized as a convex functor on the category of probability spaces and probability-preserving processes. All categorical background will be supplied.

Contents

1. Entropy and information as uncertainty—some early theorems 1
2. Other instances of entropy 4
3. The category of finite probability spaces 6
4. The convex structure on the category of finite probability spaces 9
5. Entropy is an affine functor 13
References 15

1. Entropy and information as uncertainty—some early theorems

Definition 1.1. A *finite probability space* is a pair (X, p) consisting of a finite set X and a function $p : X \longrightarrow \mathbb{R}$ satisfying

$$p(x) \geq 0 \qquad \forall\, x \in X \tag{1.2}$$

and

$$\sum_{x \in X} p(x) = 1. \tag{1.3}$$

The elements of X are called *events* and p is called a *probability distribution* on X.

Date: December 2, 2016.

Example 1.4. Let $X = \{H, T\}$ be a two-element set (H stands for "heads" and T stands for "tails") and let p be the constant function $\frac{1}{2}$. Then (X, p) is interpreted as a *fair coin toss*. Let $\lambda \in [0, 1]$ and define $p_\lambda : X \longrightarrow \mathbb{R}$ to be

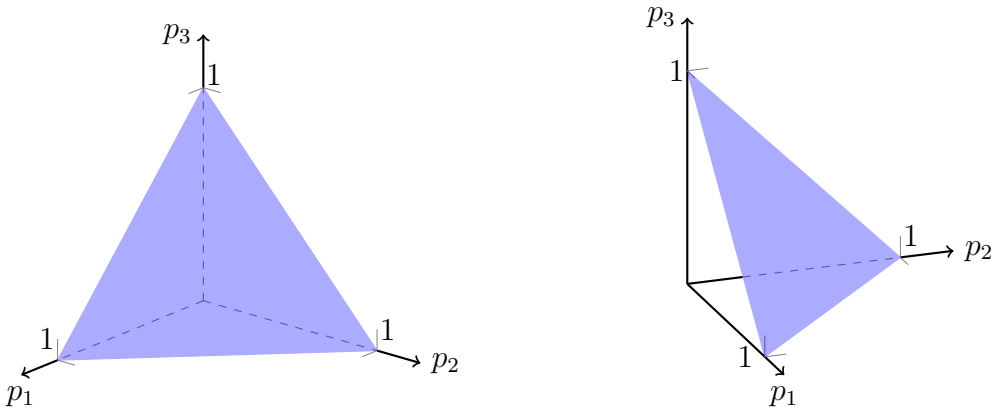
$$p_\lambda(H) := \lambda \qquad \& \qquad p_\lambda(T) := 1 - \lambda. \tag{1.5}$$

Then (X, p_λ) is interpreted as an *unfair coin toss* whenever $\lambda \neq \frac{1}{2}$. Both situations are referred to as a *coin toss*. Let **2** denote the fair coin toss.

The set of probability distributions on a finite set X of cardinality $n := |X|$ is in one-to-one correspondence with elements of an $(n-1)$-simplex Δ^{n-1}, which is defined to be

$$\Delta^{n-1} := \left\{ (p_1, \dots, p_n) \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^{n} p_i = 1 \text{ and } p_i \geq 0 \;\forall\; i = 1, \dots, n \right\}. \tag{1.6}$$

For example, the 2-simplex looks like the following subset of $\mathbb{R}^3$ viewed from two different angles



Note, also, that the permutation group S_n acts on $\mathbb{R}^n$ by permutation of coordinates and this action induces one on Δ^{n-1} for all $n \in \mathbb{N}$.

Definition 1.7. Let Δ denote the disjoint union of all simplices,

$$\Delta := \coprod_{n=1}^{\infty} \Delta^{n-1}. \tag{1.8}$$

Theorem 1.9 (Fadeev's characterization [13])**.** *Let $H : \Delta \longrightarrow \mathbb{R}$ be a function satisfying the following four axioms.*

(a) For each $n \in \mathbb{N}$, the diagram

$$\begin{array}{ccc} \Delta^{n-1} & \xrightarrow{\sigma} & \Delta^{n-1} \\ & {\scriptstyle H} \searrow \quad \swarrow {\scriptstyle H} & \\ & \mathbb{R} & \end{array} \tag{1.10}$$

commutes for all $\sigma \in S_n$.

(b) The function $H : \Delta^1 \longrightarrow \mathbb{R}$ *is continuous.*
(c) $H(\frac{1}{2}, \frac{1}{2}) = 1$.
(d) For each $n \in \mathbb{N}$,

$$H\big(tp_1, (1-t)p_1, p_2, \ldots, p_n\big) = H(p_1, \ldots, p_n) + p_1 H(t, 1-t) \tag{1.11}$$

for all $t \in [0, 1]$.

Then

$$H(p_1, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log_2 p_k. \tag{1.12}$$

Physics 1.13. The interpretation of the function H is the measure of uncertainty or the information content of a probability distribution. Fadeev's axioms can therefore be interpreted in the following way.

(a) Relabeling equally probable events does not change the uncertainty of the system.
(b) Infinitesimally perturbing the unfair coin toss results in an infinitesimal perturbation of the uncertainty.
(c) This is a normalization to specify the elementary unit for measuring uncertainty or information. The amount of uncertainty or information in the fair coin toss is declared to be 1.
(d) Consider a set $X = \{x_1, \ldots, x_n\}$ of n events with probabilities $p_1, \ldots, p_n$, respectively. Duplicate the event x_1 and label the two copies y_1 and z_1 so that they are distinguishable. Then, apply a coin toss of weight $\lambda \in [0, 1]$ to choose one of these two copies. The resulting set consists of $n + 1$ elements $\{y_1, z_1, x_2, \ldots, x_n\}$ and the probability distribution $\lambda p_1, (1-\lambda)p_1, p_2, \ldots, p_n$, respectively. The measure of uncertainty for this new probability distribution is the uncertainty in the first probability distribution plus the uncertainty of the coin toss multiplied by the probability of the event x_1 of occurring.

One can show from these axioms that the maximal measure of uncertainty on n elements occurs for the probability distribution $p_k = \frac{1}{n}$ for all $k \in \{1, \ldots, n\}$.

Definition 1.14. The function H from Fadeev's Theorem is known as the *Shannon entropy* [14]. If (X, p) is a finite probability space, then

$$H(p) := -\sum_{x \in X} p(x) \log_2 \big(p(x)\big) \tag{1.15}$$

is the *Shannon entropy* of (X, p).

If you thought that last axiom from Fadeev's Theorem was difficult to understand, then you might appreciate more intuitive yet mathematically faithful axiomatizations for the Shannon entropy, i.e. the measure of uncertainty of a probability distribution. Describing one such axiomatization is our goal.

2. Other instances of entropy

The notion of entropy appears in several different contexts yet shares many similar features. Probably the most ubiquitous property that is common to all forms is concavity (which I will refer to as convexity throughout this discussion). This is because entropy is often interpreted as the lack of information or ignorance. And if you take two systems and combine them in a convex fashion, then you lose the information that distinguishes from which of the subsystems a state comes from [16]. Thus, one expects the ignorance to be larger than the weighted sum of the individual entropies.

Example 2.1. In infinite probability theory, the entropy of a probability density function on a probability space can be defined. Given a measure space $(X, \mathcal{M}, \mu)$ consisting of a set X, a sigma algebra $\mathcal{M}$, and a measure μ satisfying $\mu(X) = 1$, together with a measurable function $p : X \longrightarrow \mathbb{R}$ satisfying $p(x) \geq 0$ for all $x \in X$ and $\int_X p(x) d\mu(x) = 1$, the entropy is defined to be

$$H\big((X, \mathcal{M}, \mu, p)\big) := -\int_X p(x) \ln\big(p(x)\big) d\mu(x). \tag{2.2}$$

Example 2.3. In quantum mechanics of a finite number of particles, the state of a system is described by a density matrix ρ, a trace-class Hermitian operator on a separable Hilbert space $\mathcal{H}$. The *von Neumann entropy* of ρ is defined to be

$$H_{\text{vN}}(\rho) = -\text{tr}\big(\rho \ln \rho\big), \tag{2.4}$$

where one uses the functional calculus for operators to define $\ln \rho$. The von Neumann entropy is a measure of the entanglement of a state. However, it is not a faithful measure of the entanglement in the sense that "$H_{\text{vN}}(\rho) > 0$ if and only if ρ is entangled" is not true [10].

d to verify this last
ment.

Example 2.5. More generally, such as in some quantum mechanical systems of an infinite number of particles or for some quantum field theories, the state of a system is described by a normal state on a von Neumann algebra [12] (or even more generally by a state on some C^*-algebra [15]). The entropy of such a state can be defined, though the formula is a bit cumbersome to explain here and will therefore be omitted. The definition is due to Narnhofer and Thirring with earlier work by Araki [9], [1]. Note that $\mathcal{B}(\mathcal{H})$ is a von Neumann algebra and there is a one-to-one correspondence between normal states on $\mathcal{B}(\mathcal{H})$ and density matrices [7], and the formula for this entropy reduces to von Neumann's formula.

Example 2.6. There are several notions of dynamical entropies for dynamical systems, which are usually described by some mathematical object together with a self-map. For example, a metric space (X, d) together with a uniformly continuous map $T : X \longrightarrow X$ has associated with it a type of entropy whose formula is somewhat complicated and will therefore be omitted. These and closely related notions of entropy were developed by Andrey Kolmogorov, Yakov Sinai, Rufus Bowen, Efim Dinaburg, Roy Adler, Alan Konheim, and M. McAndrew [5].

Example 2.7. All of the previous examples are actually just special cases of *dynamical* entropy of C^*-algebras and von Neumann algebras. This was shown to be the case by Connes, Narnhofer, and Thirring in 1987 [4].

Perhaps the most surprising definitions of entropy, therefore, arise outside the study of dynamical systems.

Example 2.8. In Lorentzian geometry, the study of manifolds with Lorentzian metrics, some of the solutions to Einstein's equations for the metric produce black hole solutions (my apologies, but I'm actually a bit ignorant about what this means precisely so I can't give a mathematical definition here). Stephen Hawking and Demetrious Christodoulou in 1970 discovered that under any physical process, the total area of all black holes in a system never decreases, in the theory of classical Lorentzian geometry. In 1972, Jacob Bekenstein, motivated by the relationship to the second law of thermodynamics, argued that the entropy of a black hole should be proportional to its surface area [3]. However, it was not clear at all what this entropy had to do with the entropy of Gibbs or Boltzman. In 1974, Hawking studied the fluctuations of quantum fields and showed that the area of a black hole *can* decrease due to the spontaneous emission of particles, imagined as coming from an infinitesimal neighborhood of the horizon [8]. Furthermore, the entropy of this emitted radiation is proportional to the area of the black hole and the precise formula is given by

$$S_{\text{BH}} = \frac{k_B A c^3}{4G\hbar}, \tag{2.9}$$

where k_B is Boltzmann's constant, G is Newton's gravitational constant, c is the speed of light, and $\hbar$ is Planck's constant. This formula can be more concisely explained in terms of the Planck length, which is defined to be

$$\ell_{\text{P}} = \sqrt{\frac{G\hbar}{c^3}}. \tag{2.10}$$

In this case, the entropy of a black hole is given by

$$S_{\text{BH}} = \frac{k_B}{4}\left(\frac{A}{\ell_{\text{P}}^2}\right). \tag{2.11}$$

In other words, ignoring Boltzmann's constant, the entropy of a black hole is proportional to one-fourth the the area of a black hole as measured using Planck units. Just to give you an idea for how huge this number is, the entropy of a black hole of radius 1 milimeter is

$$S_{\text{BH}}(R_{\text{BH}} = 1 \text{ mm}) = \frac{k_B}{4}\left(\frac{4\pi \times 10^{-9}}{3\ell_{\text{P}}^2}\right) = 5.6 \times 10^{37} \tag{2.12}$$

Compare this to the thermodynamic entropy of a liter of water, which is

$$S_{\text{H}_2\text{O liquid}} = 3.9 \times 10^3. \tag{2.13}$$

Therefore, the amount of water needed to have the same entropy as that of a black hole of radius 1 milimeter is

$$\frac{5.6 \times 10^{37}}{3.9 \times 10^{3}} = 1.4 \times 10^{34} \tag{2.14}$$

liters. The radius of a spherical body of water needed to hold this many liters of water is

$$r = \sqrt[3]{\frac{3 \times 1.4 \times 10^{34}}{4\pi \times 10^{3}}} = 1.5 \times 10^{10} \tag{2.15}$$

meters. Compare this to the distance from the Sun to Earth which is 1.5×10^{11} meters.

In all of these examples, the entropy is often defined by some formula. Occasionally (though to the best of my knowledge, not always), there are theorems, such as Fadeev's theorem, characterizing the different types of entropies. However, all these theorems and the formulas for the entropy depend on the precise details of the mathematics surrounding them. Is there a framework that describes all such instances without referring to their technical details? Is there a general over-all perspective that says what entropy actually *is* rather than just provide a *formula* in all such instances? I do not know the answer to this question, but my personal perspective is that category theory, which disassociates itself from the details of any particular theory in mathematics, might provide some insight into these questions.

3. The category of finite probability spaces

Definition 3.1. Let (X, p) and (X', p') be two finite probability spaces. A *probability-preserving* function from (X, p) to (X', p') is a function $f : X \longrightarrow X'$ satisfying

$$p'(x') = \sum_{x \in f^{-1}(x')} p(x) \tag{3.2}$$

for all $x' \in X'$.

Example 3.3. The meaning of a probability-preserving function between probability spaces can be seen nicely in the following example. Consider the set

$$X := \left\{ \boxed{\bullet}, \boxed{\begin{smallmatrix}\bullet & \\ & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix} & \bullet \\ \bullet & \end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}}, \boxed{\begin{smallmatrix}\bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet\end{smallmatrix}} \right\} \tag{3.4}$$

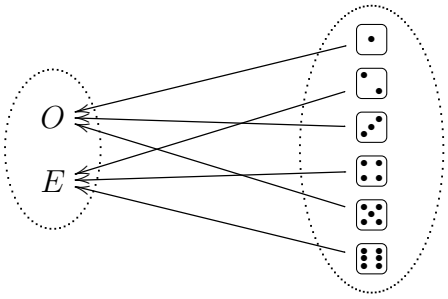
with the probability distribution given by $\frac{1}{6}$ for each element. Consider the set

$$X' := \{O, E\} \tag{3.5}$$

consisting of just two elements (O stands for "odd" and E stands for "even") with probability distribution given by $\frac{1}{2}$ for each element. Also consider the function f :

$X \longrightarrow X'$ defined by

(3.6)



sending ⚀, ⚂, and ⚅ to O and sending ⚁, ⚃, and ⚄ to E. Therefore, the probability-preserving function f is associated to the process of rolling a die and considering the likelihood of rolling an odd or an even roll versus the question of rolling a 1, 2, 3, 4, 5, or 6. There is information loss in this process because you have replaced the finer structure of the exact numbers with a coarser invariant, namely the parity of the number—odd or even. We want to quantify this information loss, and so we will seek reasonable postulates to do so.

In this regard, we must look at the *category* of finite probability spaces and probability preserving functions. The reason to emphasize category theory here is because category theory is the abstract study of processes, and if one would like to understand all kinds of entropy in terms of information change subject to some process, category theory might provide a suitable framework for such investigations.

Definition 3.7. A *category* $\mathcal{C}$ consists of a class $\mathcal{C}_0$, elements of which are referred to as *objects*, a set $\mathcal{C}_1(b,a)$, elements of which are referred to as *morphisms* from a to b, for any pair of objects a and b, a function $\mathcal{C}_1(c,b) \times \mathcal{C}_1(b,a) \longrightarrow \mathcal{C}_1(c,a)$, called the *composition*, for every triple of objects a, b, and c, and a morphism $\mathrm{id}_a \in \mathcal{C}_1(a,a)$, called the *identity* at a, for every object a. These data are subject to several conditions.

Definition 3.8. Let **FinProb** be the category specified by the following data. $\mathbf{FinProb}_0$ is the class of finite probability spaces (X,p). Given two probability spaces (X,p) and (X',p'), $\mathbf{FinProb}_1\big((X',p'),(X,p)\big)$ is the set of probability-preserving functions from (X,p) to (X',p'). The composition is the composition of functions. The identity at (X,p) is the identity function.

Remark 3.9. Technically, we should define $\mathbf{FinProb}_1\big((X',p'),(X,p)\big)$ to be the set of a.e. equivalence classes of probability-preserving functions from (X,p) to (X',p') since one could include events with zero probability in a probability space without altering the probability distribution. This technicality is actually very important but will be ignored to a large extent for the purposes of keeping this presentation more accessible.

The Shannon entropy $H : \Delta \longrightarrow \mathbb{R}$ extends to what is known as a *functor*.

Definition 3.10. Let $\mathcal{C}$ and $\mathcal{D}$ be two categories. A *functor* $F : \mathcal{C} \longrightarrow \mathcal{D}$ consists of two functions $F_i : \mathcal{C}_i \longrightarrow \mathcal{D}_i$ for $i = 1, 2$ satisfying the following two conditions.

i) For all triples of objects a, b, c in $\mathcal{C}$ and for all pairs of composable morphisms $c \xleftarrow{g} b \xleftarrow{f} a$, the diagram

$$\begin{array}{ccc} & F_0(b) & \\ {\scriptstyle F_1(g)}\swarrow & & \nwarrow{\scriptstyle F_1(f)} \\ F_0(c) & \xleftarrow[F_1(g\circ f)]{} & F_0(a) \end{array} \tag{3.11}$$

commutes, i.e. $F_1(g \circ f) = F_1(g) \circ F_1(f)$.

ii) For all objects a in $\mathcal{C}$,

$$F(\mathrm{id}_a) = \mathrm{id}_{F(a)}. \tag{3.12}$$

Example 3.13. Let A be any set. From A, one can construct a category, denoted by the same letter, specified by the following data. The set of objects is given by the set A itself. The set of morphisms from a to a' is the empty set if $a \neq a'$ and is a single element set if $a = a'$. In the latter case, this element is denoted by id_a. The composition is uniquely specified. This is called the *discrete category* associated to A. A function $f : A \longrightarrow B$ from a set A to a set B defines a unique functor on the associated discrete categories satisfying the condition that it agrees with f on the set of objects.

Example 3.14. The Shannon entropy function defines a functor $H : \mathbf{FinProb}_0 \longrightarrow \mathbb{R}$ on the associated discrete categories by setting

$$H\big((X,p)\big) := -\sum_{x\in X} p(x)\log_2\big(p(x)\big) \tag{3.15}$$

for all finite probability spaces (X, p).

This perspective is one where information content is associated to a probability space, and not a process, which is what we have been motivated to discover. This functor satisfies many familiar conditions that will be explained in more detail later. However, it is also beneficial to instead consider the following category that is a slight modification of $\mathbb{R}$.

Example 3.16. Let $\mathbb{BR}$ be the category specified by the following data. $\mathbb{BR}_0$ is a single element set, denoted by $\bullet$. $\mathbb{BR}_1(\bullet, \bullet)$ is the set $\mathbb{R}$ of real numbers. The composition is addition of real numbers. The identity at $\bullet$ is the number 0.

With this category, we can use the Shannon entropy to define a function that might potentially represent the information loss associated to probability-preserving functions.

Example 3.17. Let $F : \mathbf{FinProb} \longrightarrow \mathbb{BR}$ be the functor that sends a probability-preserving function $f : (X, p) \longrightarrow (X', p')$ to the real number[1]

$$F\Big((X',p') \xleftarrow{f} (X,p)\Big) := \sum_{x'\in X'} p'(x')\log_2\big(p'(x')\big) - \sum_{x\in X} p(x)\log_2\big(p(x)\big). \tag{3.18}$$

[1]By definition of $\mathbb{BR}$, the functor F automatically sends all probability spaces to the unique object $\bullet$ of $\mathbb{BR}$.

It is worthwhile considering special cases of this functor. First, notice that associated to every finite probability space (X,p), there is a unique probability preserving function to $\mathbf{1}$, the probability space consisting of a single element with probability distribution equal to 1. This function sends every element of X to the unique element of $\mathbf{1}$. Denote this function by $!_{(X,p)}$. Then applying the definition of F gives

$$F\Big(\mathbf{1} \xleftarrow{!_{(X,p)}} (X,p)\Big) = -\sum_{x\in X} p(x)\log_2\big(p(x)\big). \tag{3.19}$$

As an additional example, notice that for the probability preserving function f of Example 3.3 that sends ⚀, ⚁, ⚂ to O and ⚃, ⚄, ⚅ to E, this gives

$$F(f) = \sum_{k=1}^{2}\frac{1}{2}\log_2\frac{1}{2} - \sum_{k=1}^{6}\frac{1}{6}\log_2\frac{1}{6} = -\log_2 2 + \log_2 6 > 0. \tag{3.20}$$

This is actually a general phenomenon, $F\Big((X',p') \xleftarrow{f} (X,p)\Big) \geq 0$ for all probability preserving functions between any two finite probability spaces. This follows from the following calculation

$$\begin{aligned}
\sum_{x'\in X'} p'(x')\log_2 p'(x') &= \sum_{x'\in X'}\left[\left(\sum_{x\in f^{-1}(x')} p(x)\right)\log_2\left(\sum_{\overline{x}\in f^{-1}(x')} p(\overline{x})\right)\right] \\
&= \sum_{x'\in X'}\left[\sum_{x\in f^{-1}(x')} p(x)\log_2\left(\sum_{\overline{x}\in f^{-1}(x')} p(\overline{x})\right)\right] \\
&= \sum_{x'\in X'}\left[\sum_{x\in f^{-1}(x')} p(x)\log_2\left(p(x) + \sum_{\overline{x}\in f^{-1}(x')\setminus\{x\}} p(\overline{x})\right)\right] \\
&\geq \sum_{x'\in X'}\left[\sum_{x\in f^{-1}(x')} p(x)\log_2 p(x)\right] \\
&= \sum_{x\in X} p(x)\log_2 p(x)
\end{aligned} \tag{3.21}$$

since the logarithm is a monotonically increasing function. The fact that F is a functor is quite easy to see as the value on any morphism is the difference between the Shannon entropy at its source and target.

This functor, originally described by Baez, Fritz, and Leinster [2] satisfies many other wonderful properties, but to properly describe all of them, we must understand the convex structure buried within the category **FinProb**.

4. The convex structure on the category of finite probability spaces

Δ^{n-1} is an example of a convex set. Normally, one defines a convex set as a convex subspace of $\mathbb{R}^n$ for some $n \in \mathbb{N}$. However, it is useful to have an abstract definition

independent of an embedding to be able to apply such a definition in a categorical context [6].

Definition 4.1. A *convex set* is a set C together with a family of functions known as *convex linear combinations* $F_\lambda : C \times C \longrightarrow C$ indexed by $\lambda \in [0,1]$ satisfying the following axioms:

$$F_0(x,y) = y \qquad \text{(unit law)} \tag{4.2}$$

$$F_\lambda(x,x) = x \qquad \text{(idempotency)} \tag{4.3}$$

$$F_\lambda(x,y) = F_{1-\lambda}(y,x) \qquad \text{(parametric commutativity)} \tag{4.4}$$

$$F_\lambda\Big(F_\mu(x,y),z\Big) = F_{\lambda \llcorner \mu}\Big(x, F_{\lambda \lrcorner \mu}(y,z)\Big) \qquad \text{(deformed parametric associativity)} \tag{4.5}$$

for all $x,y,z \in C$ and $\lambda, \mu \in [0,1]$. Here

$$\lambda \llcorner \mu := \lambda\mu \qquad \& \qquad \lambda \lrcorner \mu := \begin{cases} \frac{\lambda(1-\mu)}{1-\lambda\mu} & \text{if } \lambda\mu \neq 1 \\ \text{arbitrary} & \text{if } \lambda = \mu = 1 \end{cases}. \tag{4.6}$$

Here "arbitrary" means that one can assign any value to the quantity.

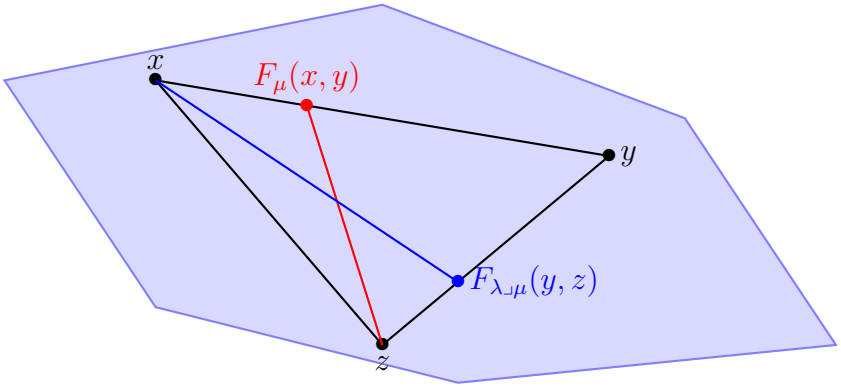
It is convenient to use the notation

$$\lambda x + (1-\lambda)y := F_\lambda(x,y). \tag{4.7}$$

In this case, the laws take a more familiar form

$$\begin{aligned} 0x + 1y &= y \\ \lambda x + (1-\lambda)x &= x \\ \lambda x + (1-\lambda)y &= (1-\lambda)y + \lambda x \\ \lambda\Big(\mu x + (1-\mu)y\Big) + (1-\lambda)z &= (\lambda \llcorner \mu)x + (1 - \lambda \llcorner \mu)\Big((\lambda \lrcorner \mu)y + (1-\lambda \lrcorner \mu)z\Big) \end{aligned}$$

The formula for deformed parametric associativity becomes significantly more apparent when one draws a picture.



The lore of category theory is to replace equations/conditions with data, and then have those data satisfy other conditions.

Definition 4.8. Let $\mathcal{C}$ be a category and let a, b be two objects in $\mathcal{C}$. An *isomorphism* from a to b, written as $b \overset{\cong}{\leftarrow} a$ is a morphism $b \overset{f}{\leftarrow} a$ for which there exists a morphism $a \overset{f^{-1}}{\longleftarrow} b$ such that the diagrams

(4.9)

both commute.

Definition 4.10. A *convex category* is a category $\mathcal{C}$ together with a family of functors known as *convex linear combinations* $F_\lambda : \mathcal{C} \times \mathcal{C} \longrightarrow \mathcal{C}$ indexed by $\lambda \in [0,1]$ and isomorphisms/morphisms

$$F_0(x,y) \overset{\cong}{\to} y \qquad \text{(unitors)} \tag{4.11}$$

$$F_\lambda(x,x) \to x \qquad \text{(idempoters)} \tag{4.12}$$

$$F_\lambda(x,y) \overset{\cong}{\to} F_{1-\lambda}(y,x) \qquad \text{(parametric commutators)} \tag{4.13}$$

$$F_\lambda\Big(F_\mu(x,y),z\Big) \overset{\cong}{\to} F_{\lambda \llcorner \mu}\Big(x, F_{\lambda \lrcorner \mu}(y,z)\Big) \qquad \text{(deformed parametric associators)} \tag{4.14}$$

for all objects $x, y, z \in \mathcal{C}$ and for all $\lambda, \mu \in [0,1]$. These data are subject to several conditions.

The "conditions" mentioned in the above theorem are referred to generically in category theory as "coherence conditions." Heuristically, they turn the picture following the definition of a convex set "fuzzy" as if there is an additional dimension. Just as for convex sets, it is convenient to use the notation

$$\lambda x \oplus (1-\lambda)y := F_\lambda(x,y) \tag{4.15}$$

for objects x, y in $\mathcal{C}$ and similarly

$$\lambda f \oplus (1-\lambda)g := F_\lambda(f,g) \tag{4.16}$$

for morphisms $x' \overset{f}{\leftarrow} x$ and $y' \overset{g}{\leftarrow} y$ in $\mathcal{C}$. We, however, use the direct sum notation to remind ourselves that we are summing objects and morphisms in some category and not just elements in some set. There are several examples of such structures, but we will stick to finite probability spaces.

Example 4.17. For every $\lambda \in [0,1]$, define the convex sum F_λ on objects by

$$\begin{gathered} \lambda(X,p) \oplus (1-\lambda)(Y,q) := \big(X \amalg Y, \lambda p \oplus (1-\lambda)q\big), \\ \text{where} \quad \big(\lambda p \oplus (1-\lambda)q\big)(z) := \begin{cases} \lambda p(z) & \text{if } z \in X \\ (1-\lambda)q(z) & \text{if } z \in Y. \end{cases} \end{gathered} \tag{4.18}$$

The convex sum of morphisms $(X', p') \xleftarrow{f} (X, p)$ and $(Y', q') \xleftarrow{g} (Y, q)$ is given by

$$\big(\lambda f \oplus (1-\lambda)g\big)(z) := \begin{cases} f(z) & \text{if } z \in X \\ g(z) & \text{if } z \in Y. \end{cases} \tag{4.19}$$

Before defining the other structure needed for a convex category, let us understand what this means in an explicit example. Let (X, p) be the probability space from Example 3.3, namely $X := \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ and $p(x) = \frac{1}{6}$ for all $x \in X$. Let $(Y, q) := (X, p)$. Fix $\lambda \in [0, 1]$. The meaning of $\lambda(X, p) \oplus (1-\lambda)(Y, q)$ is to be interpreted as the probability space for which the events in X are distinguished from the events in Y in some way and the events in X have an additional weight of λ attached while the events in Y have an additional weight of $1-\lambda$ attached. In other words, one first flips a coin to determine which set, X or Y, to land on and then one follows the probability distribution of the set X or Y once this action has been executed. One can therefore visualize these die as being colored

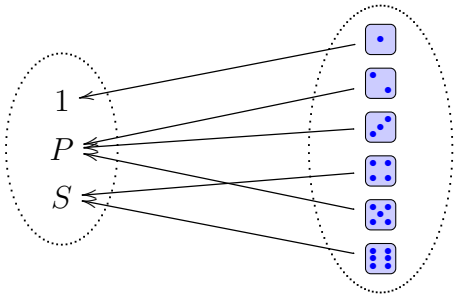
$$X \amalg Y = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}. \tag{4.20}$$

So, for example,

$$\big(\lambda p \oplus (1-\lambda)q\big)\,(\boxdot) = \frac{\lambda}{6} \qquad \text{while} \qquad \big(\lambda p \oplus (1-\lambda)q\big)\,(\boxdot) = \frac{1-\lambda}{6}. \tag{4.21}$$

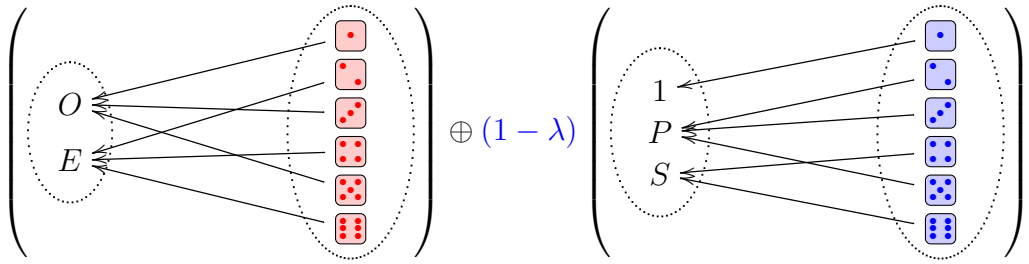
Now consider the probability preserving map $f : (X, p) \longrightarrow (X', p')$ from Example 3.3. Here $X' = \{O, E\}$ and p' is the uniform distribution of $\frac{1}{2}$. Consider the set $Y' := \{1, P, S\}$ (1 stands for 1, P stands for "prime," and S stands for "else") with probability distribution $q'(1) = \frac{1}{6}$, $q'(P) = \frac{1}{2}$, and $q'(S) = \frac{1}{3}$ together with the probability preserving function $g : (Y, q) \longrightarrow (Y', q')$ defined by

(4.22)



The function $\lambda f \oplus (1-\lambda)g$

(4.23)



can be interpreted as the possibility of one of two processes to occur where the decision of which process occurs is made by flipping a λ-coin. So far, this describes the convex

linear combinations of objects and morphisms in **FinProb**. The other isomorphisms and morphisms from the definition of a convex category are left out for brevity, and we will let the reader think about what they should be (they are somewhat obvious).

Several crucial remarks are in order in regards to this example and the general structure of convex categories.

Remark 4.24.

i) It is not true that $(X,p)\oplus(Y,q)$ is a finite probability space whenever (X,p) and (Y,q) are because the net probability of $(X,p)\oplus(Y,q)$ would be 2. This is the reason for requiring convex linear combinations.

ii) Using probability-preserving functions as the morphisms, there is no isomorphism $0(X,p)\oplus 1(Y,q)\longrightarrow(Y,q)$ for arbitrary finite probability spaces (X,p) and (Y,q). This is because a function cannot be an isomorphism if the cardinality of its domain does not equal the cardinality of its codomain. This is why technically the morphisms of **FinProb** should be taken to be a.e. equivalence classes of probability-preserving functions so that events with probability zero do not change the probability space.

iii) There is in general no reasonable probability-preserving function $(X,p)\longrightarrow\lambda(X,p)\oplus(1-\lambda)(X,p)$. This is why one only asks for a morphism $\lambda(X,p)\oplus(1-\lambda)(X,p)\longrightarrow(X,p)$ as opposed to an isomorphism in the definition of a convex category. Situations for which this occurs do exist, but not for all examples of interest.

Example 4.25. Every convex set C provides a natural convex structure on the discrete category associated to C. All the morphisms/isomorphisms are identities.

Example 4.26. Let $\mathbb{BR}$ be the one-object category whose morphisms are all real numbers with composition addition. Then the convex sum on objects is trivial (because there is only a single object) and the convex sum on morphisms (which are real numbers) is defined by

$$F_\lambda(a,b) := \lambda a + (1-\lambda)b \tag{4.27}$$

for all $\lambda\in[0,1]$ and all $a,b\in\mathbb{R}$. The other morphisms in the data of a convex category are set to be the identity (which is the number 0 in this case). In fact, $\mathbb{BR}$ is a "cone category," a concept closely related to the notion of a convex category. Essentially, this means that all objects and morphisms in the category can be scaled by any non-negative number.

5. Entropy is an affine functor

Definition 5.1. A functor $F:\mathcal{C}\longrightarrow\mathcal{D}$ is *affine* if for every $\lambda\in[0,1]$,

$$F\big(\lambda x\oplus(1-\lambda)y\big) = \lambda F(x)\oplus(1-\lambda)F(y) \tag{5.2}$$

for all objects x,y in $\mathcal{C}$ and

$$F\big(\lambda f\oplus(1-\lambda)g\big) = \lambda F(f)\oplus(1-\lambda)F(g) \tag{5.3}$$

for all morphisms f and g in $\mathcal{C}$.

Remark 5.4. Just as there are linear, convex, and concave functions on convex sets, there are even more possibilities for functors between convex categories. In particular, the subject of convex analysis is a special case of the study of convex categories and the different types of convex functors between them. This is discussed further in my thesis [11]. For the purposes of this review, we only need to focus on convex affine functors.

Example 5.5. The functor $F : \mathbf{FinProb} \longrightarrow \mathbb{BR}$ defined by

$$F\Big((X', p') \xleftarrow{f} (X, p)\Big) := \sum_{x' \in X'} p'(x') \log_2 \big(p'(x')\big) - \sum_{x \in X} p(x) \log_2 \big(p(x)\big) \tag{5.6}$$

is affine.

Remark 5.7. Compare this result to the Shannon entropy! The Shannon entropy formula is *not* affine with respect to convex combinations of finite probability spaces! It is strictly *convex*. The deviation from the Shannon entropy being affine is precisely cancelled when one takes the difference as is done in the definition of $F : \mathbf{FinProb} \longrightarrow \mathbb{BR}$.

Definition 5.8. A sequence of morphisms

$$(X_n, p_n) \xrightarrow{f_n} (Y_n, q_n) \tag{5.9}$$

in **FinProb** converges to a morphism $(X, p) \xrightarrow{f} (Y, q)$ if

(a) there exists an $N \in \mathbb{N}$ for which $X_n = X$, $Y_n = Y$, and $f_n = f$ for all $n \geq N$ and
(b) both

$$\lim_{n \to \infty} p_n = p \qquad \& \qquad \lim_{n \to \infty} q_n = q, \tag{5.10}$$

where this limit is taken in the spaces,

$$\Delta^{|X|-1} \qquad \& \qquad \Delta^{|Y|-1}, \tag{5.11}$$

respectively.

A functor $K : \mathbf{FinProb} \longrightarrow \mathbb{BR}$ is continuous if

$$\lim_{n \to \infty} K(f_n) = K(f) \tag{5.12}$$

whenever $\{f_n\}$ is a sequence in **FinProb** converging to f.

Example 5.13. The functor $F : \mathbf{FinProb} \longrightarrow \mathbb{BR}$, the difference of the Shannon entropies between the source and target, is continuous.

Theorem 5.14 (Baez-Fritz-Leinster [2])**.** *Let $G : \mathbf{FinProb} \longrightarrow \mathbb{BR}$ be a functor satisfying the following conditions.*

i) $G(f) \geq 0$ *for all morphisms* f *in* **FinProb**.
ii) G *is continuous.*
iii) $G\big(\mathbf{1} \xleftarrow{!_{\mathbf{2}}} \mathbf{2}\big) = 1$.
iv) G *is affine.*

Then $G = F$.

The first condition associates a non-negative number to every process. G is to be interpreted as a measure of the information loss associated to a process. As we discussed above in the examples, this is what we concluded for the functor F. Functoriality reflects the fact that the information loss associated an n-step process is the sum of the information losses over all n steps. Continuity says that the information loss is perturbed in a continuous fashion when the probability spaces and processes are perturbed infinitesimally. $G\big(\mathbf{1} \xleftarrow{!_{\mathbf{2}}} \mathbf{2}\big) = 1$ says that the information content of the fair coin toss is 1. Finally, the fact that G is affine says something quite intuitive. If you have two processes f and g and you toss a λ-coin to decide on which of the two processes will occur, then the information loss associated to this probabilistic process $\lambda f \oplus (1-\lambda)g$ is the λ-convex combination of the information loss for each process.

References

[1] Huzihiro Araki. Relative entropy of states of von Neumann algebras. *Publ. Res. Inst. Math. Sci.*, 11(3):809–833, 1975/76.

[2] John C. Baez, Tobias Fritz, and Tom Leinster. A characterization of entropy in terms of information loss. *Entropy*, 13(11):1945–1957, 2011.

[3] Jacob D. Bekenstein. Black holes and entropy. *Phys. Rev. D (3)*, 7:2333–2346, 1973.

[4] A. Connes, H. Narnhofer, and W. Thirring. Dynamical entropy of C^* algebras and von Neumann algebras. *Comm. Math. Phys.*, 112(4):691–719, 1987.

[5] Tomasz Downarowicz. *Entropy in dynamical systems*, volume 18 of *New Mathematical Monographs*. Cambridge University Press, Cambridge, 2011.

[6] Joe Flood. Semiconvex geometry. *J. Austral. Math. Soc. Ser. A*, 30(4):496–510, 1980/81.

[7] Brian C. Hall. *Quantum theory for mathematicians*, volume 267 of *Graduate Texts in Mathematics*. Springer, New York, 2013.

[8] S. W. Hawking. Particle creation by black holes. *Comm. Math. Phys.*, 43(3):199–220, 1975.

[9] Narnhofer Heide and Walter Thirring. From relative entropy to entropy. *FZKAAA*, 17:257–265, 1985.

[10] Elliot H. Lieb. Topics in quantum entropy and entanglement, 2014. Lectures during Princeton Summer School for Condensed Matter Physics (PSSCMP).

[11] Arthur Parzygnat. Some 2-categorical aspects in physics, 2016. Ph.D. Thesis *CUNY Academic Works*.

[12] Miklós Rédei and Stephen Jeffrey Summers. Quantum probability theory. *Stud. Hist. Philos. Sci. B Stud. Hist. Philos. Modern Phys.*, 38(2):390–417, 2007.

[13] Alfréd Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I*, pages 547–561. Univ. California Press, Berkeley, Calif., 1961.

[14] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423, 623–656, 1948.

[15] Robert M. Wald. *Quantum field theory in curved spacetime and black hole thermodynamics*. Chicago Lectures in Physics. University of Chicago Press, Chicago, IL, 1994.

[16] Alfred Wehrl. General properties of entropy. *Rev. Modern Phys.*, 50(2):221–260, 1978.

Mathematics Department, University of Connecticut, Storrs, CT 06269, USA, *Email:* arthur.parzygnat@uconn.edu